# 有关美国金融监管机构合规执法数据的语义网络分析

## 项目简介

### 项目背景

这是 Fidelity 人工智能孵化器自然语言处理部门的一项数据分析实地项目。

我从 06/2020 开始从事这个项目。该项目预计将在 2020 年 8 月下旬结束。

富达投资（Fidelity）在高度监管的环境中为客户提供服务；与美国联邦和州一级的监管机构建立规则，调查潜在违规行为，处以罚款并通过查询请求信息。为了确保公司合规，富达合规团队分析对其他公司的罚款，更新了内部政策以确保遵守新的或更新的规则并响应查询请求。目前，这项工作是被动的，手动的和费时的。

在该实地项目中，布兰代斯大学团队与富达投资的合规团队合作，通过数据分析和商业智能方式更好地理解监管机构、法规、公司、个人、违规以及施加的罚款之间的联系。

### 项目目标

1. 根据 SEC 过去 7 年有关执法行动的文件，建立有关监管者、法规、公司、个人、罚款和违法行为等 6 个节点的静态语义网络以及交互式语义网络。
2. 应用语义网络显示的关系来识别罚款的模式和趋势，并以此为基础建立预测模型。

### 项目成果

1. 使用 python 将有 SEC （美国证监会）有关执法行动的文件（xml 格式）转换为 csv 文件。（共 1038 个文件，33370 条执法记录）
[code] [sample data]
（原始 xml 文件是 Fidelity 的非公开文件。此处仅显示示例数据，包含 xml 文件的原始结构（节点名称未改变）。每个节点中的内容已被替换或简化。）

2. 在 MYSQL 中建立关系型数据库来存储执法行动数据。
[ER diagram]

3. 在 MYSQL 中执行查询操作，得到关于执法数据的一系列基本模式或规律；并从数据库中提取有关罚款或违规的执法数据。

4. 使用 python 对每个执法记录的摘要提供一系列标签，包括被罚款或违反特定规则的**公司/个人的名称**以及每个摘要中提到的**罚款金额**。
[code]

5. 建立有关监管者、法规、公司、个人、罚款和违法行为等 6 个节点的静态语义网络以及交互式语义网络。